A High Speed Sequence Search Method for Next Generation Sequencers

Sanchit Misra[1], Ramanathan Narayanan[1], Simon Lin[2], Alok Choudhary[1]
[1]Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA
[2]Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

1 Motivation

- Dream of Personalized medicine
  - Finding cause of a patient's disease in his/her DNA
  - Selecting which drug to prescribe taking into account probabilities of obtaining the desired results and experiencing side effects
  - Knowing in advance which diseases a person has a higher risk of getting
- Sequence based measurements of RNA molecules in a cell rather than the traditional method of gene expression microarrays
- Identifying polymorphisms between two sequences
  - Insertions and deletions
  - Copy number variants
  - Single nucleotide polymorphisms (SNPs)

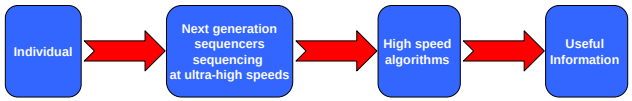
(http://davegoblog.files.wordpress.com/2007/09/genome.jpg)

2 Problem Definition

Recent advances in Genome sequencing technology have led to affordable desktop-sized machines, called Next Generation Sequencers (NGSs), with low running costs and high throughputs. It is speculated that a genome of over 100 million base pairs will be sequenced in approximately 7.5 hours, with running costs as low as $1000. Hence we need fast algorithms to extract knowledge from such a rich set of data. An important challenge posed by the deluge of data from NGSs is the problem of searching for near-exact and global matches of a query sequence Q, given a canonical genomic database G. The query sequence Q is a fragment of genome generated by sequencers with length ranging from 25 to 500 bases depending on the underlying sequencing technology. The following example shows a query found in the database with 1 mismatch and 1 gap:

Database : ……TGGAGAAACCC……AGTGAGCCGAGA ……

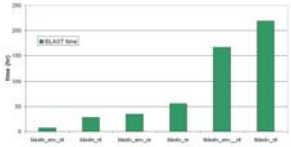
Query : GGAGATACCC……AGTGA–CCG



Picture: An example of Next generation Sequencer (http://www.454.com/)

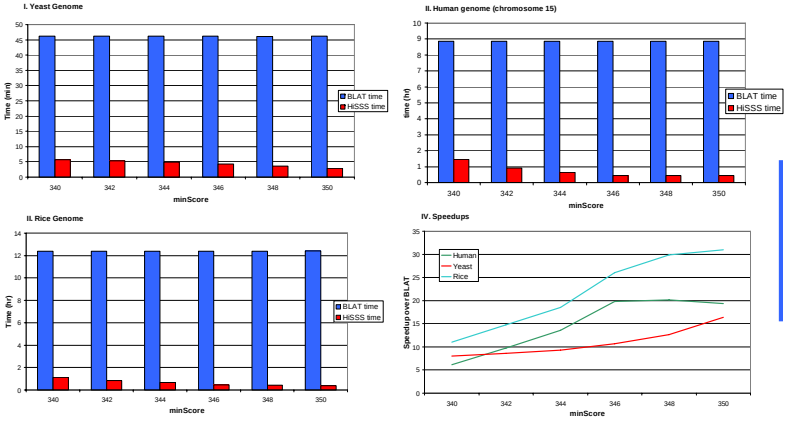
3 Why a new sequence search tool?

BLAST has been the most popular tool for sequence search and sequence alignment. We have projected the time taken by various blast search strategies for data generated by a Roche/454 sequencer in 7-8 hours. As shown in the following figure, BLAST can take hundreds of days for data generated in just one night.



BLAST creates an index of the query and scans linearly through the database. While another sequence search algorithm BLAT creates an index of the database and scans linearly through the query. The database is indexed only once in the beginning and then the index of the database is used for all queries. This greatly improves the efficiency of sequence search algorithm. But it is still not at par with the NGSs as demonstrated in the results. Since we are searching fragments of DNA in a database of same species, the query sequence is inherently very similar to the corresponding homologous region in the database. Hence we are only searching for regions in the database which nearly match the query sequence; i.e.; with very small number of indels or mismatches.

4 HiSSS: A High Speed Sequence Search Algorithm



Another important thing to note is that the speedup increases with the increase in database size from yeast to human chromosome 15 to rice. This is due to the fact that a larger database implies larger search space. HiSSS efficiently reduces this search space by judiciously ignoring the regions of the database which are considerably different from the query sequence. For all the genomes HiSSS takes well below 2 hours to process queries with a total of 100 million bases.

We Present **HiSSS**, a **Hi**gh **S**peed **S**equence **S**earch Algorithm, which dynamically reduces the search space by exploiting the near-exact match criteria.

We ran both BLAT and HiSSS on an AMD Opteron 2.4 GHz processor with 1 GB memory and running Redhat Enterprise Edition operating system. For a thorough test of our technique, we tested it against databases of yeast, human (chromosome 15) and rice genomes. We ran queries with individual lengths ~350 bases and a total of greater than 100 million bases against these databases for different values allowed mismatches and Indels. As BLAT does not have any way of specifying number of mismatches or Indels, we have used the minScore parameter to approximately simulate them. HiSSS outputs exactly same results as BLAT and uses very little extra memory. Figures I,II and III demonstrate the time taken by BLAT and HiSSS for databases containing yeast, human (chromosome 15) and rice genomes respectively. Figure IV exhibits the speedups obtained by HiSSS over BLAT for each of the three databases. **It can be clearly seen that our optimizations, consistently provide more than an order of magnitude speedup over BLAT without any loss of accuracy and with very little memory overhead.**

5 Conclusion and Future Work

- HiSSS achieves more than an order of magnitude speedup over state of the art sequence search algorithms without loss of accuracy and with very little memory overhead.
- HiSSS technique is even more beneficial for larger databases.
- HiSSS runs faster than the rate of production of data from the NGSs.
- As the Next Generation Sequencers evolve and get faster and faster, we will need even faster techniques for sequence search. This will be achieved by a combination of faster more efficient algorithms, parallel programming and/or hardware based implementations.
- These technological advancements will enable us to realize the dream of personal genomics and will completely change the way diseases are diagnosed. This will bring a tremendous changes in medical science.

6

Corresponding Author
Sanchit Misra
Graduate Student
EECS Department
Northwestern University
Evanston IL 60208
(smi539@eecs.northwestern.edu)

WebServer
http://blalah.ece.northwestern.edu/~dms146/Genome_Interface2.php